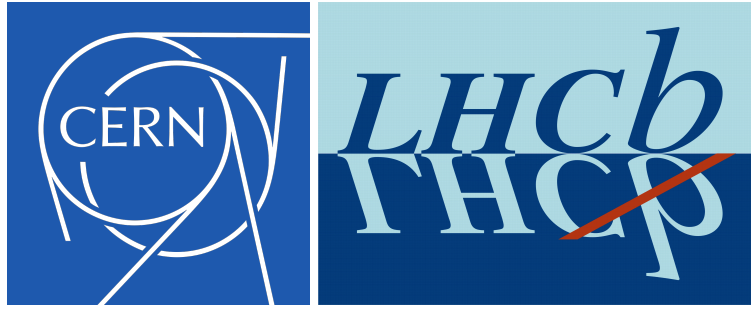




Summer Students Report

**A new approach for primary vertex
association of B -mesons using machine
learning techniques with the LHCb
experiment**

Author: Nico Schiel
E-Mail: nico.schiel@stud.uni-goettigen.de
Supervisors: Laurent Dufour
Mark Waterlaat
Start date: 30.06.25
End date: 05.09.25

Abstract

During Run 3 of the LHC, the LHCb experiment has an expected number of inelastic visible proton–proton interactions per bunch crossing of $\mu = 5.4$. Each of these interactions is called a primary vertex. Particles with a significant flight distance, such as B -mesons, must be associated with one of the reconstructed primary vertices. Currently, the association is performed by finding the primary vertex with the minimal impact parameter, defined as the distance of closest approach between the primary vertex and the B -meson track. In this project, machine learning methods with additional input features, such as the number of tracks per primary vertex and the kinematics of the B -meson, are studied as a new approach for associating primary vertices. These studies are performed on partially and fully reconstructed $B_{(s)}^{(+)}$ decays. Among the studied models, the most promising performance is obtained with a multiclass boosted decision tree that simultaneously considers the three primary vertices with the smallest impact parameters. For the partially reconstructed B -meson decay $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$, the improvement of the association efficiency is 2.0 % on average. Even larger improvements are achieved differentially for example for large number of primary vertices per event. Finally, a tool has been developed for the inclusion in the central LHCb software, which applies the model to signal candidates.

Contents

1	Introduction	1
2	The LHCb Upgrade I	1
3	Simulations, channels, object definition and event selection	2
4	Primary vertex association	3
4.1	Split and merged primary vertices	3
4.2	Impact parameter	5
4.3	Basics of machine learning	6
4.4	Feature selection	7
4.5	Binary classification	8
4.6	Multiclass classification	9
4.7	The effect of split vertices	11
4.8	Adding further input features	11
5	Incorporation into the central LHCb software	13
6	Conclusion and outlook	14
	Appendices	16
A	Separation of the input features	16
B	Efficiencies differentially	17
C	Machine learning models	18
C.1	Model parameter binary classifier	18
C.2	Model parameter multiclass classifier	18
D	Machine learning functions	19
	Bibliography	20

1 Introduction

The goal of the LHCb experiment is to study flavour physics phenomena such as CP violation and hadron spectroscopy. For these type of measurements, time-dependent and semileptonic analyses are of particular importance [1–6]. Crucial in both is the precise reconstruction of the position of the initial proton-proton collision [1–6]. During run 3 of the LHCb experiment, each bunch crossing contains on average $\mu = 5.4$ visible inelastic proton-proton collisions [7], called primary vertices (PVs). To reconstruct the flight distance, particles must be correctly associated to their true primary vertex. Otherwise, incorrect associations affect the precision of measurements. The number of PVs per bunch crossing increased compared to run 1 and 2 of the LHC and is expected to increase further with future upgrades [8]. This makes PV association increasingly challenging. Consequently, the development of improved tools for PV association is and will be essential for future analyses.

This summer student project investigates whether additional information included into a boosted decision tree or deep neural network can improve the current PV association approach. First the basics of the LHCb Upgrade I are described. The investigated decay modes and the event selection are then explained. After providing motivation for the importance of PV association, the current approach is presented. Afterwards, the newly developed methods are explained and compared both to each other and to the existing approach. In Sec. 5 a possible implementation of the newly developed methods into the central LHCb software is described. Finally, a summary and an outlook are given.

2 The LHCb Upgrade I

The LHCb Upgrade I detector is one of the four large detectors at the Large Hadron Collider (LHC) at CERN. It is the upgrade of the initial LHCb detector. The LHCb Upgrade I detector is used during the Run 3 and 4 of the LHC, which started in 2022. The instantaneous luminosity is increased up to $2 \cdot 10^{33} \text{ cm}^{-2} \text{ s}^{-1}$ with the upgrade. The layout of the detector is single-arm spectrometer. It has a coverage of $2 < \eta < 5$. It is designed to study hadrons including b and c quarks which are often produced along the beam axis [7, 9]. The detector layout is shown on the left in Fig. 2.1.

The tracking system of the LHCb Upgrade I detector is used to reconstruct the tracks of ionising particles and to measure their momenta. The first component of the tracking system is a silicon pixel detector that surrounds the proton-proton interaction region, called the Vertex Locator (VELO) [10]. The tracking system further consists of silicon-strip detectors in the upstream direction of the detector, called the upstream tracker (UT) [11]. The UT is in front of a large dipole magnet. It provides a magnetic field with a bending power of $\simeq 4 \text{ Tm}$. The magnet bends the path of electrically charged particles [11]. The magnet is followed by scintillating fibre (SiFi) trackers [11]. These measure the change of the movement direction through the magnetic field. From this change under the magnetic field the momenta of the charged particles are reconstructed [7, 9].

Two Cherenkov detectors, RICH1 and RICH2 are used for the particle identification of charged hadrons, namely of pions, kaons, and protons. RICH1 is positioned between the VELO and the UT and RICH2 behind the SiFi trackers. RICH1 and RICH2 use C_4F_{10} and CF_4 as radiators and photoelectrodes to detect the Cherenkov light [7, 9].

The RICH2 detector is followed by an electromagnetic calorimeter (ECAL) and a hadronic calorimeter (HCAL). The ECAL measures the energies of electrons and photons, and consists of lead absorbers and plastic scintillators to sample the energy deposition. The HCAL measures the energies of charged hadrons, using iron as an absorber material and scintillators [7, 9]. Photons are reconstructed using only the calorimeter information, whereas charged particles, i.e. $e^\pm$ and charged hadrons, are reconstructed using both the tracking system and the calorimeters.

Muons are not absorbed by the calorimeters since they are minimising ionising particles. Thus, the final part of the detector are the five muon stations. They are used to identify muons precisely. Each muon stations consist of alternating layers of iron and multiwire proportional chambers. The iron filters out low energetic particles, while the multiwire proportional chamber identify the muons [7, 9, 12].

Neutrinos can not be reconstructed at LHCb Upgrade I. Other particles such as short-lived hadrons like the J/Ψ are reconstructed from their decay products. [7, 9]

The tracks measured in the tracking system can be divided, depending on in which subsystem they are

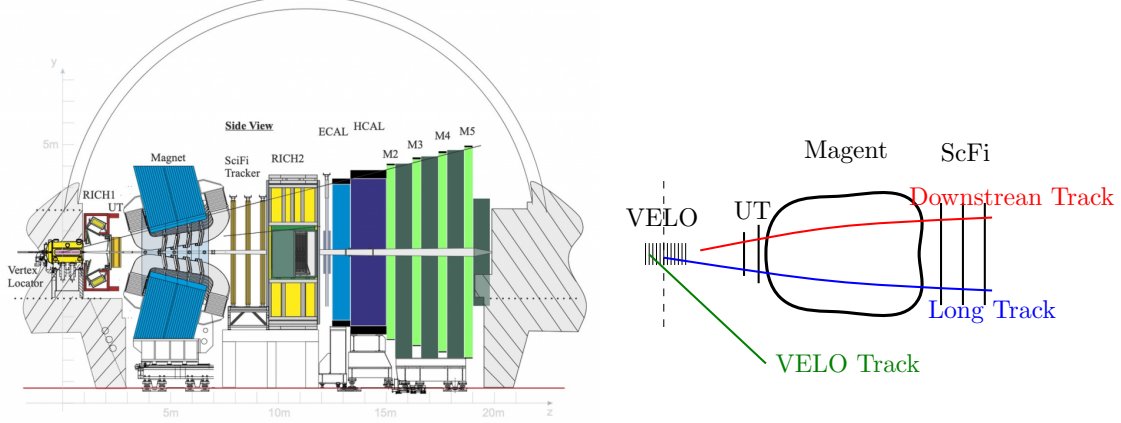


Figure 2.1: The (left) layout of the LHCb Upgrade I is shown from a side view @LHCb and the (right) different types of tracks are shown with the tracking subsystems.

Table 3.1: Mass requirements on the reconstructed hadrons in the four studied decay channels.

	$B^0 \rightarrow K^{*0} J/\Psi$	$B^0 \rightarrow K_S J/\Psi$	$B^+ \rightarrow K^+ J/\Psi$	$B_s^0 \rightarrow K^+ \mu^- \nu$
$m_{J/\Psi}$ [MeV]	[2847, 3347]	[2847, 3347]	[2847, 3347]	-
m_K [MeV]	[826, 966]	[300, 700]	-	-
m_B [MeV]	[5150, 5340]	[5150, 5340]	[5000, 5600]	[0, 5500]

measured. Tracks that are only measured in the VELO detector and not in the following subsystems are called VELO tracks. The VELO tracks have a large acceptance since also tracks in the backwards direction are measured. But no momentum reconstruction is possible. For a momentum measurement additional information from the SiFi detector are necessary. An example is tracks, which are measured in all tracking subsystems. These are called long tracks. Tracks that are measured everywhere except in the VELO, i.e. they are measured in the UT and SiFi tracker, are called downstream tracks. Downstream tracks are especially important for long-lived neutral particles like the K_s which often decay outside the VELO [7, 9, 11, 13]. Other types of tracks are not important for this report. VELO, long and downstream tracks are visualised on the right panel in Fig. 2.1.

3 Simulations, channels, object definition and event selection

The proton-proton collisions are simulated using PYTHIA [14] with a specific LHCb configuration [15]. Decays of unstable particles are modelled with EVTGEN [16] and final state radiation with PHOTOS [17]. The simulations are further corrected for resolution effects due to detector interaction with GEANT4 [18, 19], applied as described in Reference [20]. For simulated events, information that is directly available from the simulation without reconstruction are referred to as truth level. When the same events are passed through the reconstruction algorithms, the result is referred to as the reconstructed level.

To generalise the studies four different $B_{(s)}^{(+)}$ decay modes are studied. The B -mesons are reconstructed via their decay products. The final state particles are required to have a momentum of $p > 4000$ MeV and a transverse momentum of $p_T > 400$ MeV. Their pseudorapidity must be in the acceptance window of the LHCb Upgrade I detector of $2 < \eta < 5$. The decay products are truth matched to their particle type. They are also truth matched to the correct particle type of their parent particle in the simulations.

The four studied decay channels are

$$\begin{aligned}
 B^0 &\rightarrow K^{*0}(K^+\pi^-)J/\Psi(\mu^+\mu^-) \\
 B^+ &\rightarrow K^+J/\Psi(\mu^+\mu^-) \\
 B^0 &\rightarrow K_s(\pi^+\pi^-)J/\Psi(\mu^+\mu^-) \\
 B_s^0 &\rightarrow K^-\mu^+\nu_\mu
 \end{aligned}$$

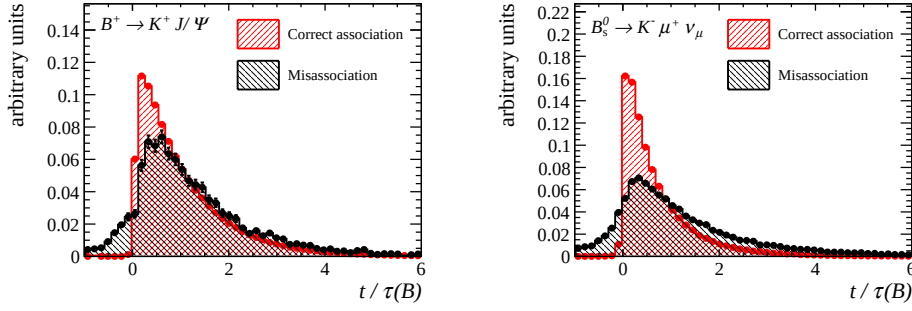


Figure 4.1: Reconstructed lifetimes normalised to the PDG value of the lifetimes of the (left) B^+ and (right) B_s^0 . The red lines are for correctly associated PVs and the black ones for misassociated PVs. The distributions are normalised.

The first three decay channels are fully reconstructed decays. Therefore, tight windows on the mass of the B -mesons [21], the J/Ψ [21] and kaons [21] are chosen, shown in Tab. 3.1. The fourth decay, $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$, is only partially reconstructed, since the neutrino is not measured. Hence, there is no lower mass cut on the B_s . Furthermore, to increase the available Monte Carlo statistics, loose requirements on the quality of the decay vertices are applied, namely $\chi^2/\text{NDF} < 16$ for the B -meson vertex and $\chi^2/\text{NDF} < 20$ for the other hadron vertices. Only long tracks are used for all reconstructed final particles, except for the pions in the $B^0 \rightarrow K_s(\pi^+\pi^-)J/\Psi(\mu^+\mu^-)$ channel. For them both long and downstream tracks are used.

4 Primary vertex association

A proton-proton collision during a bunch crossing is called primary vertex (PV). During one bunch crossing at the LHCb experiment, there are about $\mu = 5.4$ visible inelastic PVs. If a particle decays at a distance to the PV, the vertex of the decay is the secondary vertex. The secondary vertex and thus the decaying particle have to be associated to one of the PVs. PV misassociation negatively affects different types of measurements. For example time dependent measurements and the reconstruction of partially reconstructed decays depend on the correct PV association.

In more detail the proper time of a decay is calculated by measuring the distance between the PV and the SV vertex, the distance of flight d_{Flight} of the B -meson. The decay time of the particle is then defined as [22]

$$t[\text{s}] = \frac{1}{c[\text{ms}^{-1}]} \frac{d_{\text{Flight}}[\text{m}]m[\text{GeV}]}{p[\text{GeV}]},$$

where c is the speed of light, m the mass, and p the momentum of the B -meson. The reconstructed lifetimes can thus be negative due to finite vertex resolution and misassociation. That misassociation of the PV has a large effect on lifetime measurements is shown in Fig. 4.1 for $B^+ \rightarrow K^+ J/\Psi$ and $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. The distribution of the misassociated PVs is broadened compared to the correct associated PVs and the maximum is shifted. Thus, correct PV association is essential for high-precision time-dependent measurements.

4.1 Split and merged primary vertices

A split PV refers to the case where two reconstructed vertices originate from the same simulated PV. Further merged PVs are PVs where two PVs on truth level are reconstructed as one reconstructed PV.

Averaged over all samples $8.987 \pm 0.003\%$ of events contain split primary vertices (PVs). In $2.721 \pm 0.003\%$ of all events, the PV associated with the B -meson is split. In addition, $3.571 \pm 0.003\%$ of events contain a merged PV involving the B -meson. Therefore, both split and merged PVs have a non-negligible portion of events.

The PV displacement in z is defined as the absolute difference between the reconstructed PV position,

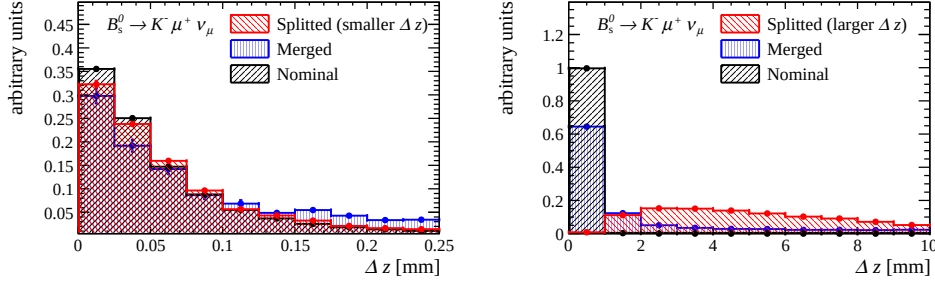


Figure 4.2: The PV displacement is shown (left) for close differences and (right) for far differences for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. It is shown for merged PVs and non-split, non-merged (nominal) PVs in both panels. The left panel shows only the reconstructed PV with the smaller Δz , whereas the right panel shows the one with the larger Δz . The distributions are normalised.

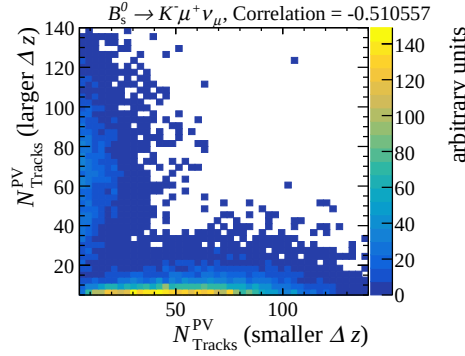


Figure 4.3: Track multiplicity correlation between the two reconstructed primary vertices (PVs) for events in which the true PV is split for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. The PVs are ordered by their distance Δz to the simulated PV, with the abscissa showing the track multiplicity of the closer PV and the ordinate that of the farther PV.

z_{reco} , and the corresponding truth level PV position, z_{truth} ,

$$\Delta z = |z_{\text{truth}} - z_{\text{reco}}|.$$

For split vertices, two Δz values are obtained, one for each reconstructed PV. One corresponding to the reconstructed PV closer to the true PV position (smaller Δz), and one corresponding to the PV with larger Δz . These two cases are shown separately in Fig. 4.2. Both figures additionally show the merged and the non-split, non-merged (nominal) PV distribution. For merged vertices, the Δz distribution is comparable to that of nominal PVs, with the only difference being a larger fraction of events at large Δz . These are expected to impact the PV association. The distribution for the closer split PV shows no significant deviation to the nominal distribution. In contrast, the distribution of the split PV with larger Δz has a strong deviation with larger values for Δz and a shifted maximum, thus having a negative impact on the PV association.

The distribution of the number of tracks per reconstructed PV in split PVs, $N_{\text{Tracks}}^{\text{PV}}$, is studied in Fig. 4.3. In most cases, one of the PVs contains significantly more tracks, while the other PV has fewer. Typically, the reconstructed PV that is closer to the simulated PV position contains the larger number of tracks. However, a non-negligible fraction of events is observed in which the reconstructed PV with the larger Δz has the higher track multiplicity.

Data contains both split and merged PVs, since there is no truth information to filter them. Therefore, the studies and results in the main part of this report are performed with split vertices included except if otherwise stated. Merged vertices are not included since the truth information for the association is not available in the studied samples.

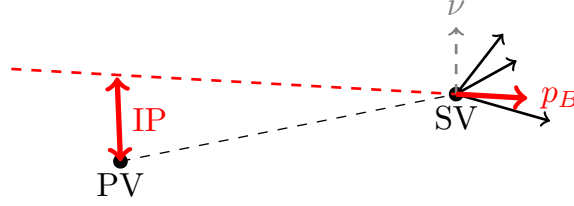


Figure 4.4: Visualisation of the decay of a semileptonic B -meson decay.

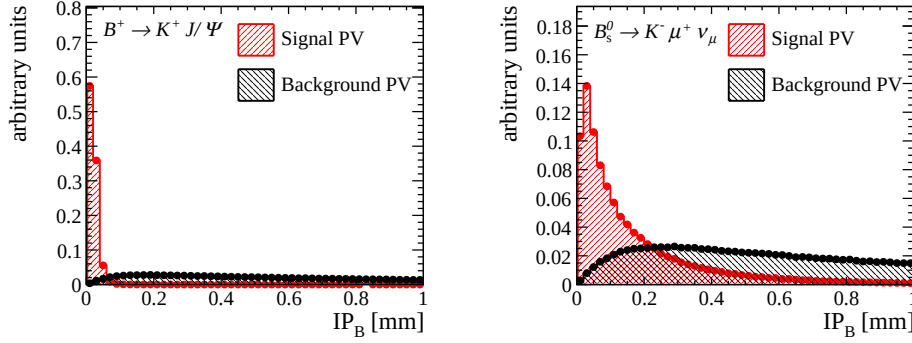


Figure 4.5: Separation of the IP for the true PV of the B -meson in red and for the two background PVs with the smallest IP in black are shown for (left) $B^+ \rightarrow K^+ J/\Psi(\mu^+ \mu^-)$ and (right) $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. The distributions are normalised.

4.2 Impact parameter

Currently, LHCb performs the PV association by taking the PV with the minimal impact parameter (IP) in an event. The IP for a particle decay is the distances of closest approach between a reconstructed PV and the reconstructed particle track, given by the position of the SV and the momentum $\vec{p}$ of the particle. Figure 4.4 visualises the definition of the IP. The IP is calculated with

$$\text{IP} = \frac{|(\vec{SV} - \vec{PV}) \times \vec{p}|}{|\vec{p}|}. \quad (4.1)$$

For the association, the IP is calculated for each PV. The PV with the smallest IP is chosen as the PV associated to the B -meson [22].

To further study the IP the separation for correct and misassociated PVs for the IP is shown in Fig. 4.5 for $B^+ \rightarrow K^+ J/\Psi$ and $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. For the background distribution only the two background PVs with the smallest IP are shown. A clear difference between the signal PV and background PV is visible for both processes. In the partially reconstructed decay $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$, the overlap is larger. The reason is that the missing momentum of the neutrino is not reconstructed. Thus, the momentum of the B -meson, which is used to calculate the IP in Equ. (4.1), is not fully reconstructed.

The association efficiency of correctly associating the PV is defined as the ratio of the number of correctly associated B -mesons $N_B^{\text{corr. asso.}}$ in a sample and the total number of B -mesons N_B in the same sample

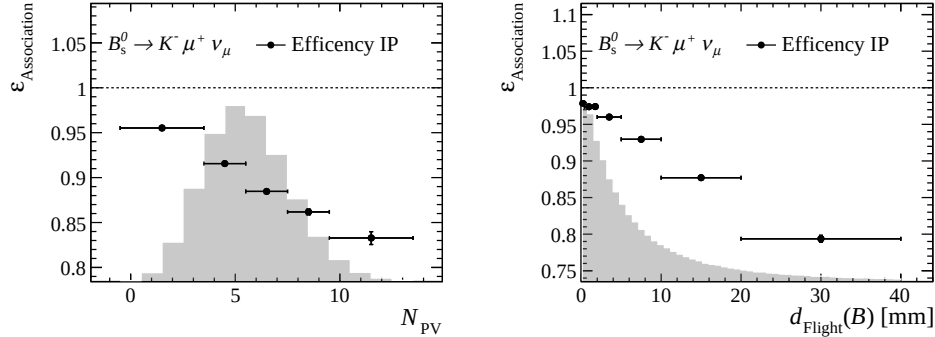
$$\varepsilon_{\text{Association}} = \frac{N_B^{\text{corr. asso.}}}{N_B}.$$

The efficiencies of the four studied processes are shown in Tab. 4.1 The efficiency for the fully reconstructed decays is high at over 99%. For the studied partially reconstructed $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ decay, the association efficiency is smaller with only $90.23 \pm 0.10\%$. The efficiency is also studied differentially in the number of PVs per event and in the flight distance in Fig. 4.6. The efficiencies decrease with larger number of PV and also decrease for large flight distances.

Summarised, the IP alone is therefore a powerful parameter for PV association. Nevertheless, significant

Table 4.1: Association efficiencies $\varepsilon_{\text{Association}}$ for the PV association with the minimum IP method.

	$B^0 \rightarrow K^{*0} J/\Psi$	$B^0 \rightarrow K_S J/\Psi$	$B^+ \rightarrow K^+ J/\Psi$	$B_s^0 \rightarrow K^- \mu^+ \nu_\mu$
$\varepsilon_{\text{Association}}$ IP	$99.50 \pm 0.03 \%$	$99.34 \pm 0.07 \%$	$99.44 \pm 0.01 \%$	$90.23 \pm 0.10 \%$

**Figure 4.6:** Association efficiencies differentially for the minimum IP method for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ for (left) the number of PVs and (right) the flight distance. The distribution of the variables is shown in grey.

improvements are possible in the partially reconstructed decay modes. Furthermore, uniform efficiency distribution as a function of certain observables such as flight distance and the number of PV is preferred.

The input events are divided into three classes, minIP1, minIP2 and minIP3. These are the events where the signal PV is the one with the smallest IP, the second-smallest IP and the third-smallest IP. The distribution of these classes, and events outside any of those classes, denoted as minIP4 can be seen in Fig. 4.7. For the fully reconstructed decays less 0.1% of events are not in the three classes minIP1, minIP2 and minIP3. For the partially reconstructed $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ decay $0.98 \pm 0.01\%$ of events are not in either of these three classes. Given their negligible contribution, events outside the minIP1, minIP2, and minIP3 classes are excluded from the newly developed approach. Nevertheless, all efficiencies quoted in this report include all the PVs.

4.3 Basics of machine learning

One possibility for an improved PV association is to use multivariate analysis (MVA) techniques. Here, instead of just one parameter, multiple input features are used to do the PV association based on a more complex decision. In supervised learning, which is used in this report, the model learns from labelled inputs and uses correlations between input features, to classify an input sample. Two different types of MVA models are studied.

The first type of classifier used in this study are gradient boosted decision tree (BDT). A BDT combines the outputs of many small decision trees into a single model. Training begins with a single decision tree that serves as an initial approximation. During each iteration, a new tree is trained to correct the errors of the current model, focusing on the parts of the data that were poorly predicted. Trees added in later stages are typically assigned smaller weights, controlled by the learning rate. The quality of the predictions is quantified using a loss function that compares the model output to the true labels. The model is updated along the gradient of the loss, thereby minimising it. This process is called gradient boosting. Training continues until the loss function converges or a predefined number of iterations is reached. To control overfitting, the maximum depth of each tree and the total number of trees are limited.

Deep neural networks (DNNs) also minimise a loss function, but their structure differs from that of BDTs. A DNN consists of an input layer, one or more hidden layers and an output layer. Each layer is composed of nodes that compute the weighted sum of the outputs from the previous layer, add a bias term, and apply a non-linear activation function. The depth of the network refers to the number of layers, while the width denotes the number of nodes in a given layer. In a fully connected architecture, as applied in this report, each node in a layer is connected to all nodes in the previous layer. The nonlinear activation functions allow the network to model complex relationships. The outputs of the final layer are compared to the true labels using the model's loss function. The network parameters, i.e. weights and biases of each node, are optimised during training by minimising the loss via backpropagation [23, 24]. An optimizer

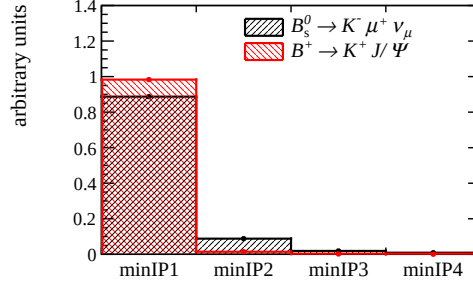


Figure 4.7: Number of Events in the classes minIP1, minIP2, minIP3 and minIP4 for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ in black and $B^+ \rightarrow K^+ J/\Psi$ in red. The distributions are normalised.

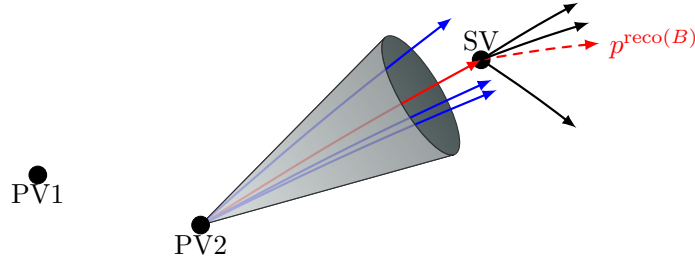


Figure 4.8: Visualisation for the cone around the movement direction of the B -meson. The central axis is shown in red as the connection of the PV and SV. Additional tracks in the cone are shown in blue and the decay products in black. The reconstructed momentum is shown as a dashed red line.

determines the update of the weights during each training step. The learning rate controls the step size. To avoid overfitting, regularisation techniques such as dropout [25] are applied.

Both BDTs and DNNs make prediction based on training data. To avoid overfitting the sample is divided into three samples. The first is the trainings sample. Based on this sample the optimisation of the model is performed. The second is the validation sample. This is used during the training to validate the results to detect and avoid overfitting. The third sample is the test sample. This is used after the training to test the performance of the model. The choice of model depends on the specific problem, data, and constraints. So both types BDT and DNN are tested. Functions that are used in the machine learning models in this report e.g. as activation function or loss are defined in App. D.

4.4 Feature selection

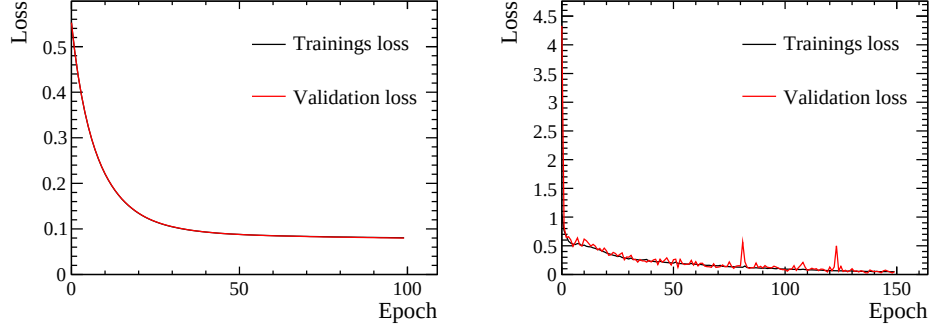
For an improved classification it is important to select input features with a significant impact. They should have a large separation between the signal PV of the B -meson and the background PVs. The final model should be as small as possible, for fast inference. Therefore, only the most sensitive features are used. Additional features that are strongly correlated with each other are avoided. The input features should also be uncorrelated to observables typical used in analyses such as the lifetime to not impact these measurements.

The IP alone is a strong discriminator as shown in Sec. 4.2. Thus, it is used as an input feature.

Additionally, variables describing the number of tracks show a strong separation between signal and background PVs. B -meson are mainly produced from a high momentum transfer q^2 parent particles. These high q^2 parent particles produce large number of additional particles, observed in the detector as tracks. Consequently, PVs of a B -hadron have on average more tracks than the background PVs. Hence, the number of tracks per PV, $N_{\text{Tracks}}^{\text{PV}}$, is an input feature for the models. Also tracks that move collinear to the B -meson are expected. To test if other tracks move collinear a cone around the movement direction of the B -meson is defined. The movement direction is not defined via the momentum of the B -meson, since this is not precisely known for partially reconstructed decays. Instead, the cone is centred around the axis that connects the SV of the B -meson and the studied PV. The definition of the cone is visualised

Table 4.2: Input features used in the binary BDT and binary DNN classification.

B -meson	PV	other PVs	Event Level
$\eta(B)$			
$p_T(B)$			
$m^{\text{visible}}(B)$			
	IP	$\Delta\text{IP 1 \& } \Delta\text{IP 2}$	
	$N_{\text{Tracks}}^{\text{PV}}$	$\Delta N_{\text{Tracks}}^{\text{PV 1 \& } \Delta N_{\text{Tracks}}^{\text{PV 2}}$	
	$N_{\text{Tracks}}^{\text{Cone}}$	$\Delta N_{\text{Tracks}}^{\text{Cone 1 \& } \Delta N_{\text{Tracks}}^{\text{Cone 2}}$	
			$N_{\text{VELO Tracks}}$
			N_{PVs}

**Figure 4.9:** Loss for the (left) binary BDT and the (right) binary DNN.

in Fig. 4.8. The opening radius is calculated as

$$(\Delta R)^2 = (\Delta\phi)^2 + (\Delta\eta)^2,$$

where $\Delta\phi$ and $\Delta\eta$ denote the differences in the azimuthal angle (in radians) and pseudorapidity between a given track and the flight direction. Optimal results are found with a large opening radius of $(\Delta R)^2 < 0.5$. The number of tracks inside this cone per PV, $N_{\text{Tracks}}^{\text{Cone}}$, is thus also used as input variable for the models. For both $N_{\text{Tracks}}^{\text{PV}}$ and $N_{\text{Tracks}}^{\text{Cone}}$ only tracks used in the reconstruction of the PV are considered. The separation for the track variables is shown in Fig. A.1 in App. A.

To relate the training to the kinematics of the b -meson, the transverse momentum $p_T(B)$, the pseudorapidity $\eta(B)$, and the visible invariant mass m^{visible} are used as input variables. The visible mass is defined as the invariant mass of the reconstructed hadrons and muons, excluding undetected particles such as neutrinos. The visible mass provides sensitivity to the missing momentum in partially reconstructed decays, for example those involving neutrinos.

Also, event level information are used as input features. The first one are the number of PVs in a bunch crossing N_{PVs} . Furthermore, the total number of VELO tracks in an event $N_{\text{VELO Tracks}}$ is used in some studied MVA models.

4.5 Binary classification

The first class of models consists of binary classifiers that consider a single PV at a time and yield the probability that this PV is the true vertex. This calculation is done for each considered PV in the event. The three PVs with minimal IP are used in the training and testing. The PV with the highest probability is chosen as the associated PV by the model. The binary classifier is trained both as a BDT and a DNN. The model parameters are given in App. C.1.

The input variables are summarised in Tab. 4.2. For a better comparison between the different PVs, also the differences in ΔF of an input feature F to the other two considered PVs are used as input. It is defined as $\Delta F = F_{\text{PV}} - F_{\text{Other PV}}$. These differences are included for all PV specific inputs, namely IP, $N_{\text{Tracks}}^{\text{PV}}$ and $N_{\text{Tracks}}^{\text{Cone}}$. The input features are min-max scaled for the DNN. For the BDT the raw values are used as input.

Table 4.3: Association efficiencies for the minimum IP method and the binary BDT and binary DNN classifier.

		$B^0 \rightarrow K^{*0} J/\Psi$	$B^0 \rightarrow K_S J/\Psi$	$B^+ \rightarrow K^+ J/\Psi$	$B_s^0 \rightarrow K^- \mu^+ \nu_\mu$
$\varepsilon_{\text{Association}}$	IP	$99.50 \pm 0.03 \%$	$99.34 \pm 0.07 \%$	$99.44 \pm 0.01 \%$	$90.23 \pm 0.10 \%$
$\varepsilon_{\text{Association}}$	BDT	$99.36 \pm 0.03 \%$	$99.31 \pm 0.08 \%$	$99.42 \pm 0.01 \%$	$91.81 \pm 0.09 \%$
$\varepsilon_{\text{Association}}$	DNN	$98.94 \pm 0.02 \%$	$99.22 \pm 0.08 \%$	$99.21 \pm 0.01 \%$	$91.83 \pm 0.09 \%$

Table 4.4: Input features used in the multiclass BDT and multiclass DNN classification.

B -meson	For each PV	Event Level
$\eta(B)$		
$p_T(B)$		
$m^{\text{visible}}(B)$		
	IP	
	$N_{\text{Tracks}}^{\text{PV}}$	
	$N_{\text{Tracks}}^{\text{Cone}}$	
		N_{PVs}

The loss for both the binary DNN and BDT converge against a common value shown in Fig. 4.9. The classification performance, quantified via the area under the ROC curve (true positive rate versus false positive rate), reaches $\text{AUC} = 1.00$ for both the BDT and the DNN, in training and validation samples.

The association efficiencies for both BDT and DNN model are shown for the four studied channels in Tab. 4.3. The machine learning models perform worse than the minimum IP method for the fully reconstructed channels. However, the partially reconstructed semileptonic $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ decay show improvement of approximately 1.6 % compared to the minimum IP method. Therefore, improvements are possible with the binary classifier in this decay channel.

4.6 Multiclass classification

The multiclass (MC) network considers 3 PVs simultaneously and associates one of those PVs to the B -meson. The goal is that this classifier gains full understanding of the event. As before, the three considered PVs are the ones with the smallest IP, i.e. the classes minIP1, minIP2 and minIP3 are used as the input vertices. However, the input PVs are not ordered by IP. Instead, the input order of the PVs is shuffled randomly. Performed studies showed that an ordering in IP results in a bias towards choosing the PV with the minimal IP more often, because of the large statistics of the minIP0 class with more than 90% of events.

For each of the 3 PVs an output is predicted, which is scaled with a softmax function. Thus, each output is the probability that the PV is the true input PV. The PV with the highest output is associated to the B -meson.

The model parameter for the BDT and DNN are summarised in App. C.2. The input features are summarised in Tab. 4.4. Min-max scaling is applied to the input features for the DNN, while the unscaled inputs are used for the BDT.

Evaluation

Both the loss of the BDT and DNN, shown in Fig. 4.10, converge against a common value. The loss of the DNN has small fluctuations even though a small learning rate of 0.00005 is used. The loss of the BDT is more stable. The AUC for both multiclass BDT and the DNN is $\text{AUC} = 0.98$ for all output variables, for the training and validation samples.

The association efficiencies for the four considered decay modes, determined using the minimum IP method, the multiclass BDT, and the multiclass DNN, are reported in Tab. 4.6. Both the DNN and BDT perform similar for the partially reconstructed decay mode and are a significant improvement on the association efficiencies. On average $\varepsilon_{\text{Association}} = 92.25 \pm 0.09\%$ are reached for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. The DNN performs worse than taking the PV with the smallest IP for the fully reconstructed decay channels. The BDT performs better than the minimum IP method for the fully reconstructed decays with

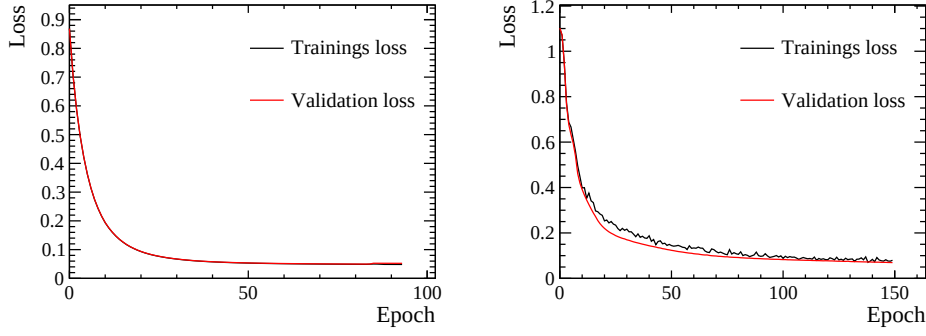


Figure 4.10: Loss for the (left) multiclass BDT and the (right) multiclass DNN.

Table 4.5: Association efficiencies for the minimum IP method and the multiclass BDT and multiclass DNN classifier.

		$B^0 \rightarrow K^{*0} J/\Psi$	$B^0 \rightarrow K_S J/\Psi$	$B^+ \rightarrow K^+ J/\Psi$	$B_s^0 \rightarrow K^- \mu^+ \nu_\mu$
$\varepsilon_{\text{Association}}$	IP	$99.50 \pm 0.03 \%$	$99.34 \pm 0.07 \%$	$99.44 \pm 0.01 \%$	$90.23 \pm 0.10 \%$
$\varepsilon_{\text{Association}}$	BDT MC	$99.26 \pm 0.02 \%$	$99.45 \pm 0.07 \%$	$99.49 \pm 0.02 \%$	$92.25 \pm 0.09 \%$
$\varepsilon_{\text{Association}}$	DNN MC	$99.15 \pm 0.03 \%$	$99.32 \pm 0.08 \%$	$99.41 \pm 0.01 \%$	$92.25 \pm 0.09 \%$

improvements up to 0.1%. The only exception is the $B^0 \rightarrow K^{*0} J/\Psi$ decay. For this the performance of the BDT is worse than the minimum IP method.

Even though the models are trained on shuffled input not ordered in IP also the confusion of the classes ordered by IP, i.e. minIP1, minIP2 and minIP3 are studied in the confusion matrices in Fig. 4.11, to better understand the classification capabilities of the models. The model is very good at classifying the events where the true PV is the one with the smallest IP. The classification for events of the other two classes is significantly worse. It only classifies about 78% (minIP2) and 75% (minIP3) correctly for the BDT. Most of the miss associated events are classified as the class minIP 1.

The feature-importance analysis, shown the multiclass BDT in Fig. 4.12, demonstrates the reason for the classifier's preference of the minIP1 class. The importance is calculated with the input features shuffled in IP. The feature importance is evaluated by replacing the values of a given input feature with random numbers drawn from a uniform distribution in the range defined by the minimum and maximum of that feature. The AUC of the classifier output is then recalculated, yielding a value denoted by AUC. Let AUC_{Nom} be the AUC obtained using the nominal, unmodified input features. The feature importance is then defined as $(\text{AUC} - \text{AUC}_{\text{Nom}})/\text{AUC}_{\text{Nom}}$. This procedure is repeated 100 times for each feature, and the average is taken. The IP of the PVs are the most important input features. Their separation power without the other variables is already very good. Hence, the classification of the minIP1 class is more precise than for the other two classes. The different track variables are also important. From these the total number of tracks is more important than the tracks inside the cone. The most important property of the B meson for the classification is the visible mass. The least important variables over all three outputs are the number of PVs and the other kinematics of the B -meson.

The association efficiencies are also studied differentially for the number of PVs per event and the flight distance d_{Flight} of the B -meson. They are shown for the minimum IP method and the multiclass BDT and DNN in Fig. 4.13 for the $B^+ \rightarrow K^+ J/\Psi$ and the $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ channel. The improvement in the association efficiencies are especially large for higher number of PVs in an event compared to the minimum IP method. For more than 10 PVs the improvements are more than 4%. The improvements of the BDT are also up to 5% for large flight distances of the B -meson for the partially reconstructed $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ decay. This is important since the long-lived component is also relevant for lifetime dependent measurements. Interesting for the $B^+ \rightarrow K^+ J/\Psi$ is, that the association efficiency decreases for small flight distances. The differential distribution of further variables of interest are shown in Fig. B.1 in App. B.

Additionally, the time for the inference of the BDT and DNN is studied. The inference should be as fast as possible if it is used in a real analysis setup. To compare the timing both models are applied via ONNX runtime in C++ [26] on a single performance thread on an Intel Ultra Core 7 155HZ. The

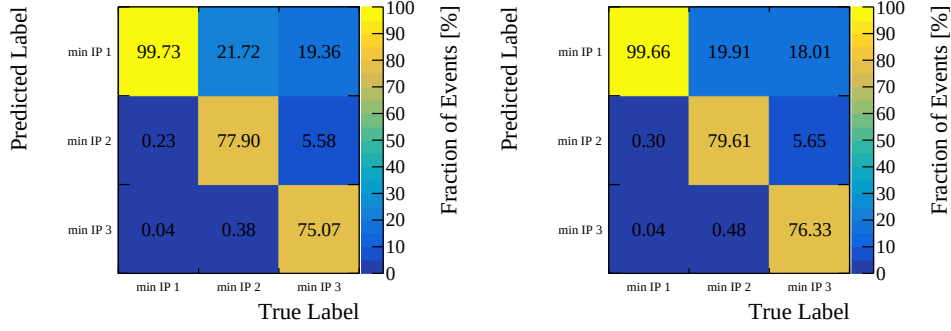


Figure 4.11: Confusion matrix for the multiclass (left) BDT and (right) DNN classifier together for all four samples for the classes minIP 1, minIP 2 and minIP 3.

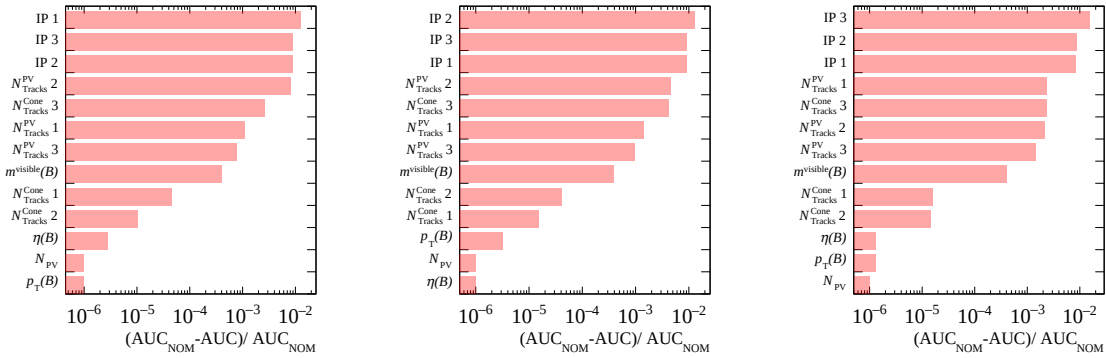


Figure 4.12: Feature importance of the multiclass BDT for shuffled inputs and for all samples together for the (left) first, (middle) second and (right) third output.

events are applied with a bunch size of one. The results for different number of events are shown in Fig. 4.14. The data points are determined by averaging about 100 tests. The behaviour is linear since a constant bunch size is used. The average inference time per event is $t_{\text{BDT}} = 0.0046 \pm 0.05 \text{ ms}$ and $t_{\text{DNN}} = 0.0071 \pm 0.04 \text{ ms}$. While the overall execution times for one event are of similar magnitude for both models, the BDT processes many events considerably faster than the DNN.

So overall the multiclass BDT model performs better than both binary classification models. It is further faster in inference than the multiclass DNN and more precise for the fully reconstructed decays. The performance for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ is the same for multiclass BDT and DNN. Therefore, based on these studies, the multiclass BDT is chosen for further studies in this report, as the currently best model.

4.7 The effect of split vertices

The studies shown in the report are all repeated with split vertices removed. That means it is both training and testing only on non-split, non-merged PVs are used. Table 4.6 shows the association efficiencies with split vertices removed. These show that all channels are improving for the multiclass BDT and DNN when split vertices are not considered. The improvement in association efficiency for the $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ is 2.4 % for only non-split, non-merged PVs. The effect is also significant for the $B^0 \rightarrow K^{*0} J/\Psi$. Without split PVs included the BDT perform better than the minimum IP method. Split primary vertices have a significant negative impact on the analysis. Several studies were carried out to identify variables capable of discriminating split PVs, but no effective method was identified.

4.8 Adding further input features

The performance of the model can be further improved by adding additional input features. One example is the distance of flight or related to this the position of the PVs and of the SV. For partially reconstructed decays another example is the corrected mass m^{corr} . The corrected mass accounts for parts of the not

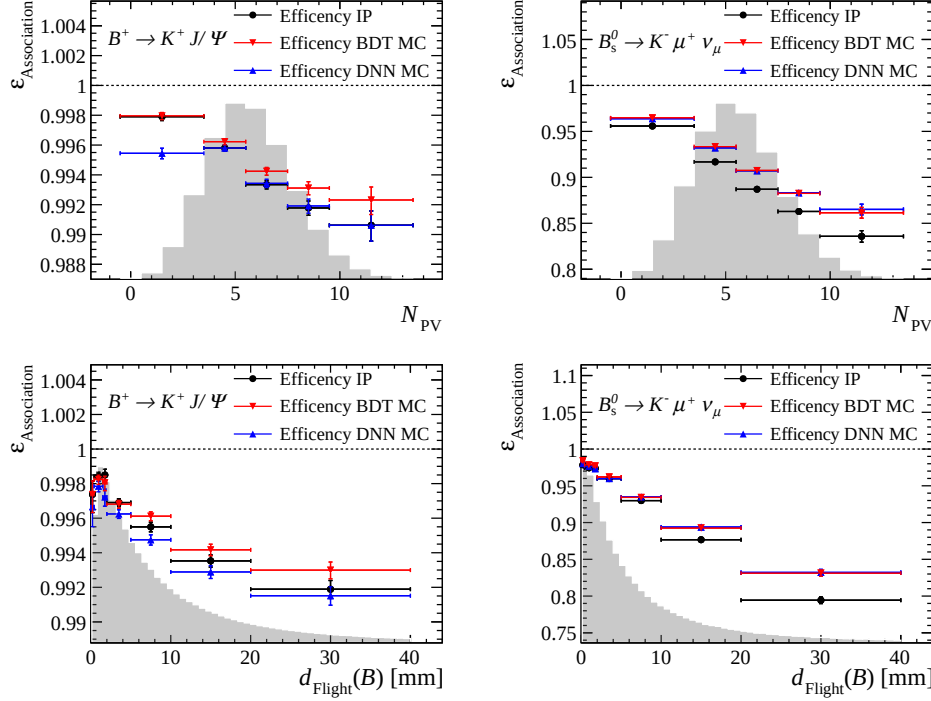


Figure 4.13: Association efficiencies differentially for the multiclass BDT in red, the multiclass DNN in blue and the minimum IP method in black for (left) $B^+ \rightarrow K^+ J/\Psi$ and (right) $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. The distribution of the variables is shown in grey. The show variables are (top) the number of primary vertices and (bottom) the flight distance.

Table 4.6: Association efficiencies for the minimum IP method and the multiclass BDT and multiclass DNN classifier with split vertices included and with split vertices excluded

		$B^0 \rightarrow K^{*0} J/\Psi$	$B^0 \rightarrow K_S J/\Psi$	$B^+ \rightarrow K^+ J/\Psi$	$B_s^0 \rightarrow K^- \mu^+ \nu_\mu$
Split vertices included					
$\epsilon_{\text{Association}}$	IP	$99.50 \pm 0.03 \%$	$99.34 \pm 0.07 \%$	$99.44 \pm 0.01 \%$	$90.23 \pm 0.10 \%$
$\epsilon_{\text{Association}}$	BDT MC	$99.26 \pm 0.02 \%$	$99.45 \pm 0.07 \%$	$99.49 \pm 0.02 \%$	$92.25 \pm 0.09 \%$
$\epsilon_{\text{Association}}$	DNN MC	$99.15 \pm 0.03 \%$	$99.32 \pm 0.08 \%$	$99.41 \pm 0.01 \%$	$92.25 \pm 0.09 \%$
Split vertices excluded					
$\epsilon_{\text{Association}}$	IP	$99.51 \pm 0.03 \%$	$99.34 \pm 0.08 \%$	$99.44 \pm 0.01 \%$	$90.17 \pm 0.10 \%$
$\epsilon_{\text{Association}}$	BDT MC	$99.61 \pm 0.02 \%$	$99.45 \pm 0.08 \%$	$99.53 \pm 0.01 \%$	$92.60 \pm 0.09 \%$
$\epsilon_{\text{Association}}$	DNN MC	$99.40 \pm 0.03 \%$	$99.03 \pm 0.08 \%$	$99.31 \pm 0.01 \%$	$92.61 \pm 0.09 \%$

reconstructed momentum e.g. from the neutrino in $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. It is defined as

$$m^{\text{corr}} = \sqrt{(m^{\text{visible}})^2 + p_{\text{T,FD}}^2 + p_{\text{T,FD}}^2}$$

where m^{visible} is the visible invariant mass introduced previously and $p_{\text{T,FD}}$ is the transverse momentum relative to the flight direction. Therefore, it can only account for the transverse and not for the longitudinal missing momentum. As shown in Fig. 4.15, for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$, the correctly associated PVs are sharply distributed at the B_s^0 mass with a skew for smaller masses. The background PV distribution is broadened. By including the corrected mass for each PV in the multiclass BDT the association efficiency $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ can be improved up to $95.21 \pm 0.08 \%$. These efficiencies are especially large at the B_s^0 mass, shown on the right of Fig. 4.15. Using a classifier that is trained on the corrected mass also favours the PVs for the background where the corrected mass is at the B -mass. Thus, this introduces a bias. This bias of the background can also affect other parts of an analysis. For example, the corrected mass is often used as a fit-parameter [5, 6]. Therefore, the usage of these input features is not advisable for analyses that use these input features. Nevertheless, if an analysis does not use these variables a

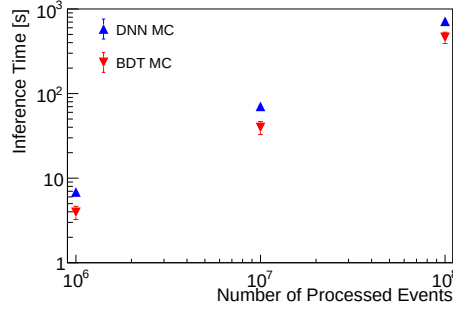


Figure 4.14: Timing of the inference of the multiclass BDT and DNN model for different number of events. Both axes are logarithmic.

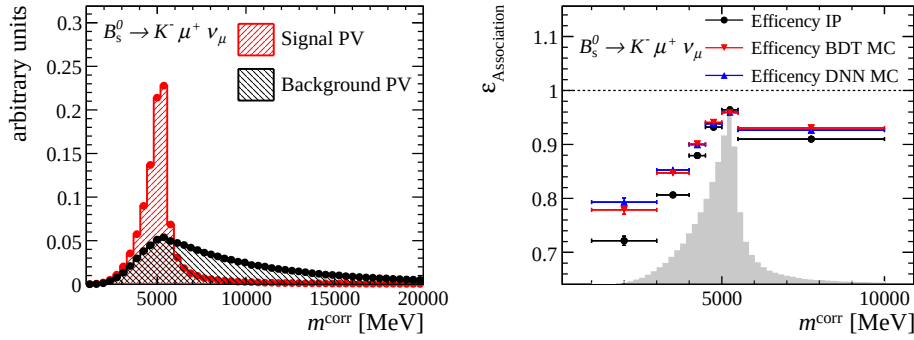


Figure 4.15: The (left) separation of the corrected mass calculated with the true PV in red and in black for the two background PVs with the smallest IP and the (right) association efficiency differentially in the corrected mass are shown. The models are trained with the corrected mass as input feature. Both figures are for $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. The distributions are normalised.

significant improvement of the PV association is possible. But since the classifier is less general the model without the corrected mass is used as baseline model in the rest of this report.

5 Incorporation into the central LHCb software

To apply the machine learning method in a real analysis it must be assessable in LHCb. For this the machine learning inference is implemented in such a way that it can be directly included into the central LHCb reconstruction analysis software [27, 28]. The developed algorithm executes successfully when integrated into the software framework but is not yet part of the central reconstruction software.

The algorithm is written as a C++ class in the LHCb reconstruction software. It takes in all the signal candidates and reconstructed PVs in an event. The algorithm associates a PV to each signal candidate according to the MVA model. Then a copy of the particle is produced and the associated PV is assigned to the copy of the signal candidate. The output of the model is a container of particles and the probability that the PV is the true PV according to the classifier.

The developed algorithm processes each event as follows. If an event contains fewer than three reconstructed PVs, no reassignment is performed. The existing signal PV stays unchanged, and a dummy value is assigned to the PV association probability. For less than three events the association is already precise. For events with at least three PVs, the algorithm first calculates the IP of the particle with respect to each PV. The algorithm then selects the three PVs with the smallest IP values as association candidates. Afterwards, all additional input variables required by the model are computed for these three PV candidates. The model is applied to the input features via ONNX Runtime [26]. By default, the multiclass BDT is applied, although a multiclass DNN implementation is also available. The ONNX model is read once during algorithm initialisation, and model inference is executed in a single thread to ensure thread safety. After the model inference, the PV with the maximum classifier output is associated to the signal candidate.

The implemented algorithm is called `PV_Association_BMesons_BDTMulticlass` and can be applied in DAVINCI. The current application in the LHCb environment for the tupling is shown in the following code example

```
from PyConf.Algorithms import PV_Association_BMesons_BDTMulticlass

def main():
    ...
    out = PV_Association_BMesons_BDTMulticlass(
        InputParticles = InputParticles,
        InputVertices  = InputVertices)
    OutputParticles    = out.OutputParticles
    OutputProbability  = out.OutputProbability
    ...
```

Performed test showed that the algorithm works as intended. It makes the same prediction as the local algorithm.

The output probability can be used to define working points for the PV association efficiency of certain processes. With a tight cut, most of the background PVs, which are wrongly associated are removed from the selection. The statistics for the correctly associated PVs are correspondingly also smaller.

6 Conclusion and outlook

This summer student project presents a new approach for PV association for B -mesons at LHCb. Instead of associating the PV with the smallest IP a multivariate analysis techniques based on machine learning is used. The additional input features are event based variables, kinematics of the B -meson and track distribution variables of each PV.

The results show that PV association can be improved with the applied machine learning techniques. Among the binary and multiclass DNN and BDT models studied, the overall best performance is achieved with the multiclass BDT. This model considers the three PVs with the smallest IP in an event and associates one of them. The improvement by this model is especially large for the partially reconstructed decay $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. An average improvement of 2.1% is achieved. Even higher improvements are achieved in certain differential regions of interest. For example the improvement for more than ten PVs in an event is over 4% in $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. Also, the improvement is larger for large flight distances. The efficiency distribution still decreases for large lifetimes, but it becomes more uniform compared to taking the PV with the smallest IP. Also, two of the three fully reconstructed decay modes show improvements compared to the minimum IP method. Only the $B^0 \rightarrow K^{*0} J/\Psi$ performs worse than the minimum IP method.

The association efficiencies are negatively impacted by the inclusion of split vertices. Without these the association efficiencies increase for the machine learning approaches. The split vertices must be included for the simulations to be comparable to data. Studies have been performed to deal better with split vertices. However, no effective method has been identified to address them.

The results are overall promising. Nevertheless, further studies need to be performed to validate the method. Firstly, a relevant study is to understand the behaviour of the $B^0 \rightarrow K^{*0} J/\Psi$ decay better and to understand why it performs worse than the other samples. Also, important is to understand split PVs better and for example find ways to identify them. In addition, the generalization of the model must be tested on decay channels with which it was not trained on. Also studies regarding the usage of machine learning in the trigger system are important. Here the main challenge is the inference time of the models. Also interesting would be to apply the model or train an additional one for Upgrade II conditions with an expected amount of $\mu = 42$ visible PVs per Event [8].

Additionally, a first implementation into the central LHCb software is developed. This implementation applies the model to a signal candidate and associates one of the PVs in the event. The algorithm could be optimised in terms of runtime by implementing it multithreaded.

Acknowledgements

I want to thank Laurent and especially Mark for their great supervision and help during this project! It was a lot of fun and a fascinating project!

I would also like to thank the LHCb group at CERN for the opportunity to do my summer student project with them and for the Moritz Karbach summer student prize. I also want to thank my fellow summer students for a great time outside of work and for a memorable summer.

Appendices

A Separation of the input features

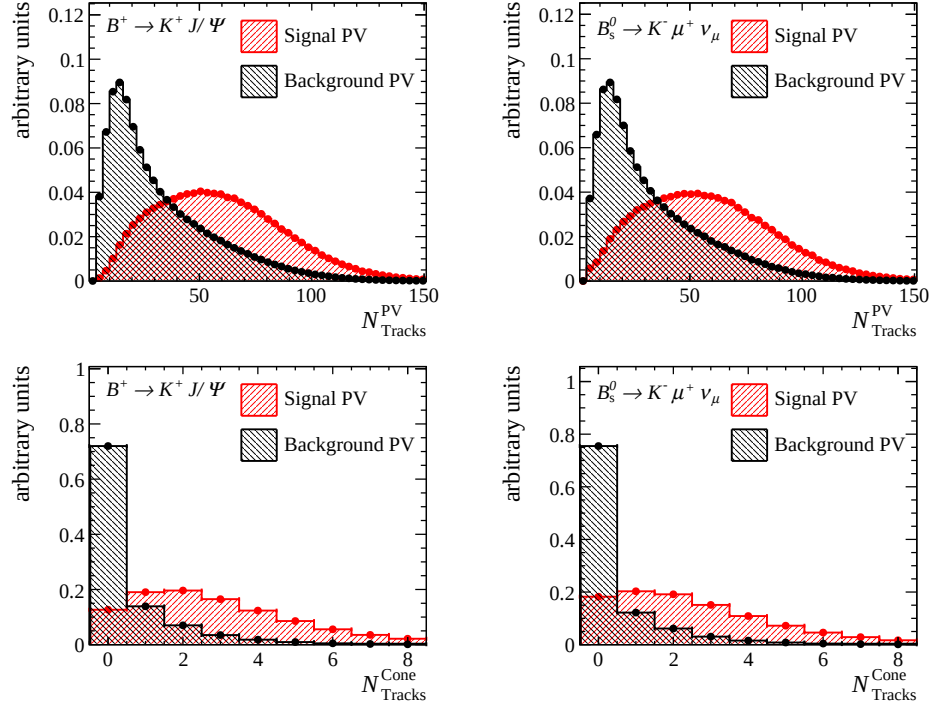


Figure A.1: Separation of $N_{\text{Tracks}}^{\text{PV}}$ and $N_{\text{Tracks}}^{\text{Cone}}$ for the true PV of the B -meson in red and for the two background PVs with the smallest IP in black for $B^+ \rightarrow K^+ J/\Psi(\mu^+ \mu^-)$ on the left and $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ on the right. At the top the number of tracks per PV is shown and at the bottom the number of tracks in the cone around the movement axis. The distributions are normalised.

B Efficiencies differentially

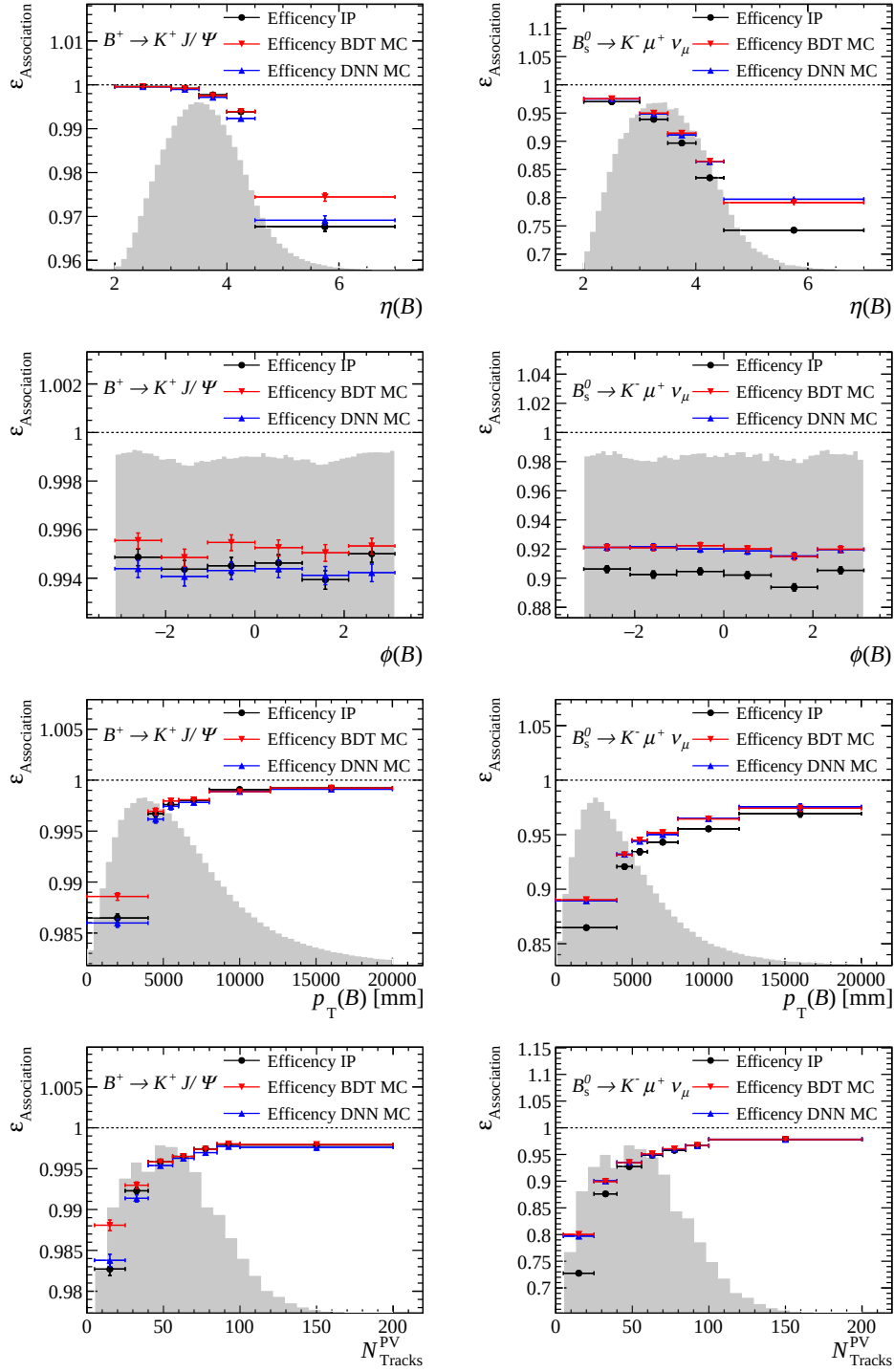


Figure B.1: Association efficiencies differentially for the multiclass BDT in red, the multiclass DNN in blue and the minimum IP method in black for (left) $B^+ \rightarrow K^+ J/\Psi$ and (right) $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$. The distribution of the variables is shown in grey. From the top to bottom the, the pseudorapidity η , the azimuthal angle ϕ and the transverse momentum p_T of the B -meson and the number of tracks of the associated PV are shown.

C Machine learning models

C.1 Model parameter binary classifier

Both BDT and DNN model are optimised for optimal performance and to avoid overfitting. The BDT has a learning rate of 0.1 and a maximal depth three per tree. The maximum number of trees is 100. The DNN consists of four layers, with 128, 64, 32 and 1 nodes. The activation functions between the layers are **ReLU** functions and the output activation function is a **sigmoid**. After the first hidden layer dropout of 10 % is applied. The optimizer used is **Adam** [29] with a learning rate of 0.00001. For both the BDT and DNN the **binary cross entropy** loss is used. The batch size is 10000.

C.2 Model parameter multiclass classifier

Both BDT and DNN model are optimised for optimal performance and to avoid overfitting. The trained BDT has a learning rate of 0.1. The maximal depth of each tree is 3 and the number of trees per class is limited by 80. For the DNN 50 different models were tested as a hyperparameter optimisation. The best performing DNN model was used. It consists of seven layers, with 256, 126, 64, 32, 16, 8 and 3 nodes. After the first layer dropout of 20 % is applied as regularisation. The activation functions between the layers are **ReLU** functions. The optimizer used is **Adam** [29] with a learning rate of 0.00005. For both the BDT and DNN the **categorical cross entropy** loss is used. The batch size is 10000. For both the model with the smallest validation loss is used.

D Machine learning functions

In this Appendix functions that are used in the machine learning models used in this report are defined.

Activation Functions

ReLU

The ReLU function for an input x is defined as

$$\text{ReLU}(x) = \max(0, x)$$

Output Activation Functions

Softmax

The `softmax` function transforms a vector of real-valued $\vec{x}$ into a set of probabilities $\vec{x}'$. It is defined as

$$x'_i = \text{softmax}(\vec{x}, i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Sigmoid

The `sigmoid` function transforms a real-valued x into a probability x' . It is defined as

$$x' = \text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

Loss Functions

Categorical Cross-Entropy

Let $\vec{t}$ be the target and $\vec{p}(\vec{x})$ be the prediction of the classifier for the input $\vec{x}$. The `categorical cross-entropy` loss function for the classification into three classes is defined as

$$\text{loss}(\vec{t}, \vec{p}(\vec{x})) = \text{CE}(\vec{t}, \vec{p}(\vec{x})) = - \sum_{i=1}^3 t_i \log(p_i(\vec{x}))$$

Binary Cross-Entropy

Let t be the target and $p(\vec{x})$ be the prediction of the classifier for the input $\vec{x}$. The `binary cross-entropy` loss function for the binary classification into two classes is defined as

$$\text{loss}(t, p(\vec{x})) = \text{BC}(t, p(\vec{x})) = -[t \log(p(\vec{x})) + (1 - t) \log(1 - p(\vec{x}))]$$

Input Scaling

Min-Max-Scaling

For one input feature f the minimax scaling for one element f_i is defined as

$$f_i^{\text{scaled}} = \frac{f_i - \min(f)}{\max(f) - \min(f)}$$

Bibliography

- [1] LHCb collaboration. “First observation of $D^0 - \bar{D}^0$ oscillations in $D^0 \rightarrow K^+ \pi^- \pi^+ \pi^-$ decays and measurement of the associated coherence parameters”. **Phys. Rev. Lett.** **116** 2016 p. 241801.
- [2] LHCb Collaboration. “Precise determination of the $B_s^0 - \bar{B}_s^0$ oscillation frequency”. **Nature Phys.** **18** 2022 pp. 1–5.
- [3] LHCb collaboration. “Measurement of $|V_{cb}|$ with $B_s^0 \rightarrow D_s^{(*)-} \mu^+ \nu_\mu$ decays”. **Phys. Rev. D** **101** 2020 p. 072004.
- [4] LHCb collaboration. “First observation of the decay $B_s^0 \rightarrow K^- \mu^+ \nu_\mu$ and Measurement of $|V_{ub}|/|V_{cb}|$ ”. **Phys. Rev. Lett.** **126** 2021 p. 081804.
- [5] LHCb Collaboration. “Determination of the quark coupling strength $|V_{ub}|$ using baryonic decays”. **Nature Phys.** **11** 2015 pp. 743–747.
- [6] LHCb Collaboration. “Search for the rare decay $B^+ \rightarrow \mu^+ \mu^- \mu^+ \nu_\mu$ ”. **Eur. Phys. J. C** **79** 2019 p. 675.
- [7] LHCb Collaboration. “The LHCb Upgrade I”. **JINST** **19** 2024 P05065.
- [8] “Future physics potential of LHCb”. 2022 LHCb-PUB-2022-012, CERN-LHCb-PUB-2022-012.
- [9] LHCb Collaboration. “The LHCb Detector at the LHC”. **JINST** **3** 2008 S08005.
- [10] LHCb Collaboration. “LHCb VELO Upgrade Technical Design Report”. 2013 CERN-LHCC-2013-021, LHCb-TDR-013, CERN-LHCC-2013-021, LHCb-TDR-013.
- [11] LHCb Collaboration. “LHCb Tracker Upgrade Technical Design Report”. 2014 CERN-LHCC-2014-001, LHCb-TDR-015, CERN-LHCC-2014-001, LHCb-TDR-015.
- [12] A. A. Alves Jr. et al. “Performance of the LHCb muon system”. **JINST** **8** 2013 P02022.
- [13] LHCb collaboration. “Study of light-meson resonances decaying to $KS^0 K\pi$ in the $B \rightarrow (KS^0 K\pi) K$ channels”. **Phys. Rev. D** **111** 2025 p. 092009.
- [14] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. “PYTHIA 6.4 Physics and Manual”. **JHEP** **05** 2006 p. 026.
- [15] I. Belyaev et al. “Handling of the generation of primary events in Gauss, the LHCb simulation framework”. **J. Phys. Conf. Ser.** **331** 2011 p. 032047.
- [16] D. J. Lange. “The EvtGen particle decay simulation package”. **Nucl. Instrum. Meth. A** **462** 2001 pp. 152–155.
- [17] N. Davidson, T. Przedzinski, and Z. Was. “PHOTOS interface in C++: Technical and Physics Documentation”. **Comput. Phys. Commun.** **199** 2016 pp. 86–101.
- [18] GEANT4 Collaboration. “GEANT4 - A Simulation Toolkit”. **Nucl. Instrum. Meth. A** **506** 2003 pp. 250–303.
- [19] GEANT4 Collaboration. “Geant4 developments and applications”. **IEEE Trans. Nucl. Sci.** **53** 2006 p. 270.
- [20] M. Clemencic et al. “The LHCb simulation application, Gauss: Design, evolution and experience”. **J. Phys. Conf. Ser.** **331** 2011 p. 032023.
- [21] S. Navas et al. “Review of particle physics”. **Phys. Rev. D** **110** 2024 p. 030001.
- [22] LHCb Collaboration. “Measurements of the B^+, B^0, B_s^0 meson and Λ_b^0 baryon lifetimes”. **JHEP** **04** 2014 p. 114.
- [23] Shun Amari. “Backpropagation and stochastic gradient descent method”. **Neurocomputing** **5** 1993.
- [24] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. **Nature** **323** 1986.
- [25] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. **J. Mach. Learn. Res.** **15** 2014.
- [26] *ONNX Runtime*. By ONNX Runtime developers. <https://onnxruntime.ai/>. Version: 1.20.00. 2021.
- [27] G. Barrand et al. “GAUDI - A software architecture and framework for building HEP data processing applications”. **Comput. Phys. Commun.** **140** 2001 pp. 45–55.
- [28] LHCb Collaboration. “Upgrade Software and Computing”. 2018 CERN-LHCC-2018-007, LHCb-TDR-017.
- [29] Timothy Dozat. “Incorporating Nesterov Momentum into Adam”. In: *Proceedings of the 4th International Conference on Learning Representations* 2016.